Enhancing predictive models for short-term forecasting electricity consumption in smart buildings

Manuel Herrera, Daniel Fosas, David A. Coley
Dept. of Architecture and Civil Engineering
University of Bath
Bath, United Kingdom
a.m.herrera.fernandez@bath.ac.uk,
d.fosas.de.pando@bath.ac.uk, d.a.coley@bath.ac.uk

Bruno M. Brentan
Centre de Recherche en Automatique de Nancy
University of Lorraine
Nancy, France
bruno.brentan@univ-lorraine.fr

***Abstract*—Lighting, heating, and air conditioning systems are instances of how electricity use at buildings is of key importance for occupants comfort and well-being. Since the electricity can be produced but cannot be stored, for utility companies it is important to reliably forecast energy supply almost in near real-time. Nowadays, smart grid technologies development also require a high resolution forecasting to eliminate blackouts and to optimally adapt energy supply to customers' needs. These are the reasons why the finest Machine Learning and Data Science based methods have been developed and applied to approach as much accurate as possible predictive models for short-term electricity consumption. This paper proposes to enhance those predictive models by using weather and calendar information to configure a more complete working database. In addition, a cluster-based forecasting methodology will augment any predictive model with learning from other buildings. Thus, predicting future values for one smart meter is approached by utilising not only its own historical electricity consumption values, but working with a multivariate time series on weather and calendar data and information from other buildings at the same cluster. This proposal has been tested with measures from smart meters collected every 30 minutes during one year for 5 selected buildings in Bristol (UK). The enhancing methodology can predict electricity consumption data with higher accuracy than using data from just one building.**

Keywords—electricity consumption; smart meters; predictive models; machine learning

I. INTRODUCTION

According to the US Energy Information Administration, the Building Sector consumes nearly half (47.6%) of all energy produced in the US and up to 75% of all the electricity produced in the country is used just to operate buildings. This figure is near 50% in Brazil and reaches 90% in the UK. This reveals the high importance of electricity both for the energy sector and the built environment. Buildings use electricity for lighting, heating, cooling, and for operating appliances and computers. So, electricity is essential for building occupants comfort and well-being where people spend nowadays 80-90% of the time [1]. Having high precision measures and accurate forecasts of domestic electricity consumption is essential for an optimal management of this resource by the side of the authorities and supply companies. In addition, smart grids will play an important role in future cities as they are supply energy networks with associated digital communications technology to detect and react to local changes in usage. Smart buildings connected to smart grids can significantly contribute to the objective of "zero carbon" for the built environment as the technology makes now possible to collect and use intermittent renewable additional energies for buildings [2]. Smart grids optimally work coupled with the so-called smart meters; electronic devices that record consumption of electric energy in intervals of an hour or less and communicates that information (at least daily) back to the utility for monitoring and billing. The European Union aims to replace at least 80% of electricity meters with smart meters by 2020 wherever it is cost-effective to do so. Only in Great Britain more than 8.6 million of smart and advanced meters operate across homes and small business

Data science combines different fields of work in statistics and computing science in order to interpret data for the purpose of decision making. Data Science models can be understood as data-driven models relying upon the presence of a considerable amount of data that is used to build models for complementing or replacing physically based models. Among others, machine learning methods (statistical techniques to provide computers the ability to "learn" from data without being explicitly programmed) represent the basis of further data-driven models development. Data science has played a key role on analysing energy use [3,4], and specifically electricity [5], during the last years. These works show how Big Data techniques and Machine Learning methods outperform traditional data analysis methods on accuracy, computational efficiency and costs. Still more, this new framework becomes nowadays a must for the smart energy management given the huge quantity of data available, hardly managed by more classical approaches. Data Science tools are also essential to meet challenges related to the current and future need of having near real-time response for the electricity consumption at buildings embedded into smart grid systems.

This paper aims at improve the way to approach predictive models for electricity consumption in smart buildings. The proposal is made up of two parts. First, the historic time series of electricity consumption database is augmented by adding weather and calendar information. This is important as it is

well known how external weather conditions directly impact in the electricity demand [6]. For instance, the use of air conditioning systems in warmer periods of the year. Calendar variables are also of main importance. Weekends, bank holidays, and other events affecting occupancy periods have a huge influence in electricity use and demand level or if there is consumption at all. There also be considered calendar dates before and after any specific holiday as the electricity use (e.g. use of home appliances for cooking or cleaning) might vary in the occupants preparation for special days. This happens in the case of occupants receive guests and also in the case of they are spending days out.

The second part proposed by this paper is to increase the database corresponding to a single building with information of other buildings of similar electricity consumption. This is made first by computing distances or similarities between whole time series of historical electricity demand. A cluster analysis based on these distances provides groups of buildings of similar characteristics. Thus, any consequent predictive model will be enhanced by using a clustering-based approach. The main advantage to work this way is to get more robust and accurate models as we use more information. In addition, this will save the further burden of computations associated to compute predictive models for each single smart meter when having a big number of homes to deal with. The challenge is to how optimally plug the information of time series data belonging to the same cluster into a common predictive system. Electricity consumption data collected every half hour from 7 smart meters placed at 5 selected buildings in Bristol (UK) is used to test the benefits of the proposed procedure.

II. MINING TIME SERIES DATABASES

This section briefly introduces indexing time series methods and how them are used for clustering (and classification of) time series databases. These techniques are part of a more general subject known as time series data mining [7].

A. Indexing time series

A time series is a set of observations, each one being recorded at a specific time. Indexing time series strategies transform the original data into index sequences. The aim is to use a reduced dimensionality representation (dimensionality of a series is related to time, and it can be understood as the length of the series) that contains most of the original information. This is useful for analysing smart meter data as those often come associated with long time series streams that require computationally efficient methods for near real-time response. Among the main time series representations proposed in the literature highlight Discrete Fourier Transform [8], Discrete Wavelet Transform [9], Piecewise Aggregate Approximation (PAA) [10], and Symbolic Aggregate approXimation (SAX) [11]. Over them, only SAX achieves a triple objective: dimensionality reduction, symbolic discretization, and the creation of lower bounding distance measures (e.g. Euclidean distance, Dynamic Time Warping). This distance is essential for further classification and clustering of whole time series.

SAX is a time series indexing process that divides the time series data into equally sized segments. This process mainly consists of time series indexing process using a variation of the PAA methods segment partition. SAX classifies PAA segments by alphabet symbols of a dictionary (e.g. {a,b,c,d}), to codify how far their mean is from the general mean of the time series. One of the key advantages using SAX is the dimensionality reduction of long time series which become into the so-called SAX words. That is, sequences of alphabet symbols (typically letters from a dictionary) which concatenate forming a SAX word. SAX introduces new metrics for measuring distance between strings by extending Euclidean and PAA distances. Some antecedents of this proposal can be found in literature. It is worthy mentioning the work of Bach et al., 2013 [12] where the authors applied indexable SAX (*i*SAX) for big data techniques development to analyse power grid time series. *i*SAX is a modification of the classical SAX through a multi-resolution data structure designed for (massive size) time series indexing and mining [13].

B. Clustering time series databases

Cluster analysis or clustering [14] is the task of grouping a set of items such that items in the same group (cluster) are more similar to each other than to those belonging to other groups. A clustering method attempts, then, to group the objects based on the definition of similarity (proximity) supplied to it. Cluster analysis plays an important role in many fields and can be used both for preliminary and descriptive data analysis and unsupervised classification, and for summarising common features of groups of elements, like identification of centroids or baricenters. For instance, Imanishi et al., 2017 [15] use the well known k-means algorithm applied to time series decomposition of power demand data to extract uncertain features.

Whole time series clustering is a type of clustering algorithm for handling dynamic data. As mentioned above, the most important element to consider is the (dis)similarity or distance measure between time series [16]. In many cases, algorithms developed for whole time series clustering take main characteristics of the static clustering algorithms. Then the process continues by modifying the similarity definition to an appropriate one or by applying a transformation to the series so that static features are obtained [17]. Therefore, the underlying basis for time-series clustering procedures remains approximately the same than those for clustering general items. The most common approaches are hierarchical and partitional clustering [16]. A different approach is to use each row time series data as elements to directly apply clustering. In this case, it is recommended a previous dimensionality reduction of the time series length to handle more suitable data. This is the approach used in this paper after a SAX dimensionality reduction of each of the time series to estimate the distances among them.

III. MINING TIME SERIES DATABASES

This paper proposes a methodology to enhance predictive models for electricity consumption in smart buildings. The main novelty is to present a predictive model based on historic

values, weather and calendar information, which is also based on time series belonging to the same cluster.

A. Time series forecasting

The main aim of time series modelling is to collect and study the past observations of a time series to develop an appropriate model which describes the inherent structure of the series. These models will also allow to predict future time series response. Time series forecasting is widely studied in numerous practical subjects such as Economics [18], Science [19] and Engineering [20]. Over the years, researchers have focused their efforts on developing efficient models to improve the forecasting accuracy.

There is a plethora of predictive methods which provide a suitable output in terms of accuracy and computational efficiency in their implementation. Among them, the classical methodology on autoregressive integrated moving average (ARIMA) and the artificial neural networks (ANNs) stand out as widely used predictive models for time series.

- ARIMA: ARIMA models are based on a linear regression that takes into account the time dependency of the process [21]. They are composed by a number of well defined elements. There is an autoregressive (AR) part from which the evolving variable of interest is regressed on its own historical values. There is also a moving average (MA) part that is a linear combination of error terms both occurred at present time and at various times in the past (lags). The series can be "integrated" (the data values are replaced with the difference between their values and previous values) to reach a constant mean of the process. This is one of the assumptions for time series stationarity and it is necessary to validate the ARIMA model. The model can count on Stational features (SARIMA) and/or with eXogenous regressor variables (ARIMAX). The purpose of each of these features is to make the model fit the data as good as possible.

- ANNs: An ANN is an interconnected group of artificial neurons. Each neuron executes a non-linear computation based on the input values and the resulting value is fed to other neurons [22]. Neurons are usually arranged as a series of interconnected layers. Based on the data presented to the network, an algorithm (usually backpropagation) is used to iteratively adjust the neuron connection weights in such a way that the predictive performance of the network is improved. The most common ANN network is the feed-forward network, which uses the backpropagation algorithm for training. Obtaining this type of network is an iterative process in which each sample case is presented several times to the input neurons of the ANN. Usually, a typical three-layer feed-forward model is used for forecasting purposes. Hidden nodes with appropriate nonlinear transfer functions are used to process the information received by the input nodes, each associated with one of the predictors.

B. Cluster-based whole time series forecasting

This paper starts with a time series dimensionality reduction based on SAX. Then, a clustering algorithm for whole time series is approached as distance measure between the time series. This serves as basis to further apply any partitional clustering method as k-means [11]. The choice of both the previous indexing step and the partitional clustering (instead of the hierarchical option) have been taken by scalability reasons. Time-series clusters has the advantage to find groups of data streams with similar response. The number of operations for approaching predictive models is reduced if there is proposed one model per cluster and not per single time series. In addition, it can automatically presents a enhanced predictive process as single time series can learn from the most similar elements in the cluster (Figure 1).

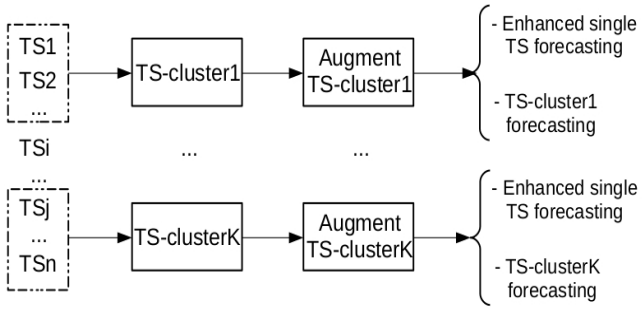

Fig. 1: Overall procedure for the proposed predictive models enhancement

Figure 1 shows an scheme of the proposed process. This can be summarised in the following steps: preprocessing single time series (by approaching a dimensionality reduction process and computing the distance between time series), clustering whole time series, augment database information with exogenous time series (calendar and weather), and decide if we focus on enhancing a single time series forecasting, as it is described above, or we propose one single predictive model per cluster (useful for the case of having a large number of clusters).

IV. Results on Smart Meters from Selected Buildings

Electricity consumption data collected every half hour from smart meters placed at 5 selected buildings in Bristol (UK) will test this paper proposal. The measures were taken during one year: 00h30m 1st January 2013 to 00h00m 1st January 2014, that is 17,520 values in total per single time series. In total, there are 7 smart meters within the 5 buildings: 3 in *Building 1* {*B1a, B1b, B1c*}, and 1 in the rest of buildings: *Building 2 - Building 5* {*B2, B3, B4, B5*}. After applying SAX based distance the 7 smart meters can be grouped in 3 clusters: *C1* = {*B1a, B1c*}, *C2* = {*B1b, B2, B3*}, *C3* = {*B4, B5*}.

In order to test this paper proposal, we compare the results obtained from using a single ANN for each building and the results coming from the enhanced ANN forecasting. For instance, working on Building *B4* the enhanced ANN encompasses weather and data information plus the measures of electricity consumption recorded at Building *B5* (same clustering). The number of hidden layer nodes remains the

same (5 nodes after tuning parameters) to propose a meaningful comparison. Still, size of training and test datasets, learning rate, and maximum number of iterations are also the same both for enhanced and single ANN. The results predicting electricity demand for every half hour during a week (336 values) confirm we gain accuracy, in the sense of the mean squared error (MSE), by using the enhanced ANN (MSE = 0.0055) with respect to the single ANN (MSE = 0.0078). Figure 2 shows that the predictions made by the enhanced ANN are (in general) more concentrated around the line $y = x$ in the plot observed vs predictions (a perfect alignment with the line would indicate a MSE of 0 and thus an ideal perfect prediction) than those made by the single ANN.

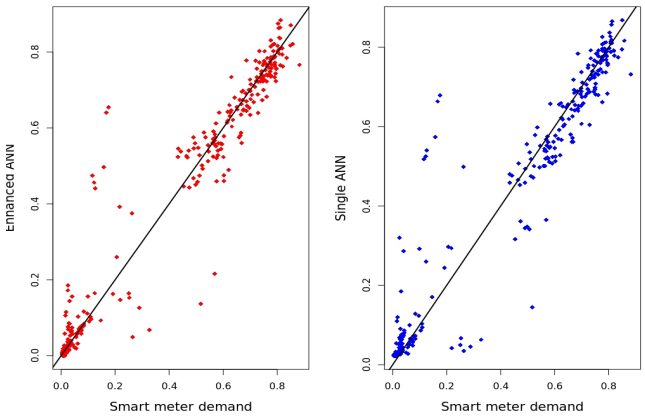

Fig. 2: Plot of electricity demand observed vs predicted for the enhanced ANN (left side, red) and the single ANN (right side, blue).

Figure 2 also shows how the enhanced ANN provides better results for lower values of the observed electricity demand. These good results come again for higher values of electricity where the single ANN has a greater tendency to underestimate the electricity demand than the predictions made by the enhanced ANN.

V. Conclusions

This paper proposes a novel framework for enhancing predictive models for short-term forecasting electricity consumption in smart buildings. Augmenting a single time series data with exogenous information provides a more accurate (multivariate) model at any case. For instance, adding extra weather and calendar information it is beneficial for any electricity consumption model in smart buildings. In addition, having a time series database is even possible to approach a clustering process of whole time series gaining further insight for modelling and forecasting. The advantages are the following:

- Imputation of time series values in case of smart sensors failures. Time series belonging to the same clustering support tackling this process in near real-time.
- Detecting correlation between time-series. This aid to understand electricity use for different buildings and occupants characteristics.
- Combining clustering and function approximation per cluster. This finds new time series patterns in addition to ease handling big size databases.
- Pattern discovery: to discover the interesting patterns in databases. For instance different patterns of electricity consumption under similar weather and calendar conditions can be discovered.

Further work come from considering clustering within the time series in addition to the whole time series groups . This will benefit of better anomaly detection to provide suitable adaptation to near real-time challenges. Exploring variations for training ANN models and different architectures, including those representing deep learning, are also avenues of further research.

Acknowledgement

This research has been performed in the project COLBE on The creation of localized current and future weather for the built environment, Engineering and Physical Sciences Research Council (EPSRC) [grant code EP/K002724/1].

References

[1] Klepeis, N.E., Nelson, W.C., Ott, W.R., Robinson, J.P., Tsang, A.M., Switzer, P., Behar, J.V., Hern, S.C. and Engelmann, W.H., 2001. The National Human Activity Pattern Survey (NHAPS): a resource for assessing exposure to environmental pollutants. *Journal of Exposure Science and Environmental Epidemiology*, 11(3), p.231.

[2] Wurtz, F. and Delinchant, B., 2017. “Smart buildings” integrated in “smart grids”: A key challenge for the energy transition by using physical models and optimization with a “human-in-the-loop” approach. *Comptes Rendus Physique*, 18(7-8), pp.428-444.

[3] Zhou, K., Fu, C. and Yang, S., 2016. Big data driven smart energy management: From big data to big insights. *Renewable and Sustainable Energy Reviews*, *56*, pp.215-225.

[4] Jain, R.K., Smith, K.M., Culligan, P.J. and Taylor, J.E., 2014. Forecasting energy consumption of multi-family residential buildings using support vector regression: Investigating the impact of temporal and spatial monitoring granularity on performance accuracy. *Applied Energy*, *123*, pp.168-178.

[5] Edwards, R.E., New, J. and Parker, L.E., 2012. Predicting future hourly residential consumption: A machine learning case study. *Energy and Buildings*, *49*, pp.591-603.

[6] [Staffell, I. and Pfenninger, S., 2018. The increasing impact of weather on electricity supply and demand. *Energy*, 145, pp.65-78.

[7] Fu, T.C., 2011. A review on time series data mining. *Engineering Applications of Artificial Intelligence*, *24*(1), pp.164-181.

[8] Chen, M.Y. and Chen, B.T., 2014. Online fuzzy time series analysis based on entropy discretization and a Fast Fourier Transform. *Applied Soft Computing*, *14*, pp.156-166.

[9] Chaovalit, P., Gangopadhyay, A., Karabatis, G. and Chen, Z., 2011. Discrete wavelet transform-based time series analysis and mining. *ACM Computing Surveys (CSUR)*, *43*(2), p.6.

[10] Chakrabarti, K., Keogh, E., Mehrotra, S. and Pazzani, M., 2002. Locally adaptive dimensionality reduction for indexing large time series databases. *ACM Transactions on Database Systems (TODS)*, *27*(2), pp.188-228.

[11] Lin, J., Keogh, E., Wei, L. and Lonardi, S., 2007. Experiencing SAX: a novel symbolic representation of time series. *Data Mining and knowledge discovery*, *15*(2), pp.107-144.

[12] Bach, F., Çakmak, H.K., Maass, H. and Kuehnapfel, U., 2013, February. Power grid time series data analysis with Pig on a hadoop cluster compared to multi core systems. In Parallel, Distributed and Network-Based Processing (PDP), 2013 21st Euromicro International Conference on(pp. 208-212). IEEE.

[13] Shieh, J. and Keogh, E., 2008, August. i SAX: indexing and mining terabyte sized time series. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 623-631). ACM.

[14] Everitt, B.S., 1979. Unresolved problems in cluster analysis. *Biometrics*, pp.169-181.

[15] Imanishi, T., Yoshida, M., Wijekoon, J. and Nishi, H., 2017, June. Time-series decomposition of power demand data to extract uncertain features. In Industrial Electronics (ISIE), 2017 IEEE 26th International Symposium on (pp. 1535-1540). IEEE.

[16] Aghabozorgi, S., Shirkhorshidi, A.S. and Wah, T.Y., 2015. Time-series clustering–A decade review. *Information Systems*, *53*, pp.16-38.

[17] Liao, T.W., 2005. Clustering of time series data—a survey. *Pattern recognition*, *38*(11), pp.1857-1874.

[18] [Granger, C.W.J. and Newbold, P., 2014. *Forecasting economic time series*. Academic Press.

[19] Bhaskaran, K., Gasparrini, A., Hajat, S., Smeeth, L. and Armstrong, B., 2013. Time series regression studies in environmental epidemiology. *International journal of epidemiology*, *42*(4), pp.1187-1195.

[20] Herrera, M., Izquierdo, J., Pérez-García, R and Ayala-Cabrera, D., 2014. On-line Learning of Predictive Kernel Models for Urban Water Demand in a Smart City. *Procedia Engineering*, *70*, pp.791-799.

[21] Peña, D., Tiao, G.C. and Tsay, R.S., 2011. A course in time series analysis (Vol. 322). John Wiley & Sons.

[22] Bishop, C.M., 1995. Neural networks for pattern recognition. Oxford university press.